Skywork UniPic 3.0: Unified Multi-Image Composition with Sequence Modeling

Skywork Multimodality Team
multimodal@skywork.ai

Project Page: https://unipic-v3.github.io

Abstract

The recent surge in popularity of Nano-Banana and Seedream 4.0 underscores the community's strong interest in multi-image composition tasks. Compared to single-image editing, multi-image composition presents significantly greater challenges in terms of consistency and quality, yet existing models have not disclosed specific methodological details for achieving high-quality fusion. Through statistical analysis, we identify Human-Object Interaction (HOI) as the most sought-after category by the community. We therefore systematically analyze and implement a state-of-the-art solution for multi-image composition with a primary focus on HOI tasks. We present Skywork UniPic 3.0, a unified multimodal framework that integrates text-to-image (T2I) generation, single-image editing, and multi-image composition. Our model supports an arbitrary number and resolution of input images, as well as arbitrary output resolutions (within a total pixel budget of 1024×1024). To address the challenges of multi-image composition, we design a comprehensive data collection, filtering, and synthesis pipeline, achieving strong performance with only 700k high-quality training samples. Furthermore, we adopt a novel training paradigm that formulates multi-image composition as sequence modeling—a conditional generation task on a unified sequence. To accelerate inference, we integrate trajectory mapping and distribution matching into the post-training stage, enabling the model to produce high-fidelity samples in just $1\tilde{}4$ steps and achieve a $50\times$ speedup over standard diffusion sampling. Skywork UniPic 3.0 achieves state-of-the-art performance on single-image editing benchmarks (GEditBench-EN: xxx; ImgEdit-Bench: xxx) and surpasses both Nano-Banana and Seedream 4.0 on multi-image composition using our newly proposed MultiCom-Bench, thereby validating the effectiveness of our data pipeline and training paradigm.

1 Introduction

The rapid advancement of diffusion models has revolutionized generative AI, enabling unprecedented capabilities in text-to-image synthesis and image editing. While early efforts primarily focused on generating images from scratch or modifying single images, recent community interest has shifted toward more complex scenarios involving multiple images. The viral success of systems like Nano-Banana [4] and Seedream 4.0 [16] demonstrates a growing demand for multi-image composition—where elements from different source images are seamlessly blended into a coherent, high-quality output.

Despite this surge in interest, multi-image composition remains fundamentally challenging. Unlike single-image editing where structural consistency is largely preserved, composing multiple images requires reconciling potentially conflicting semantics, lighting conditions, perspectives, and artistic styles. According to our statistical analysis, the community shows particular interest in the category

Tech report. Preprint.

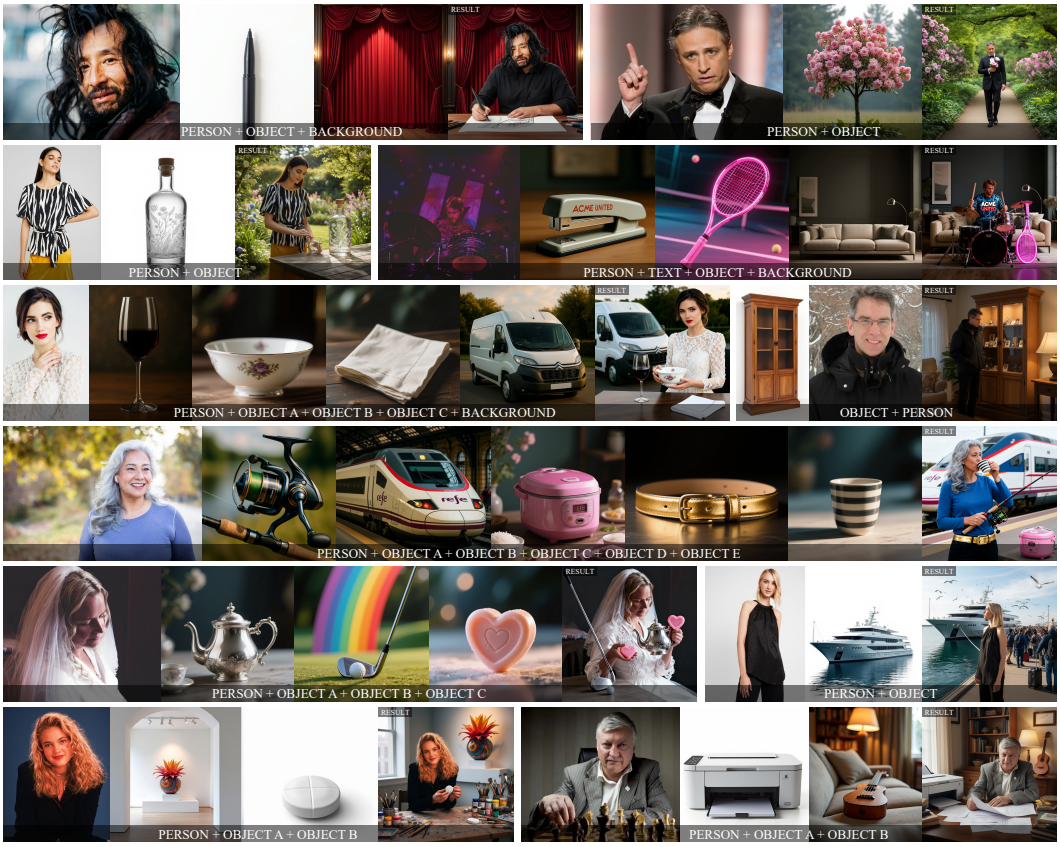

Figure 1: Our model supports image editing and composition conditioned on 1 to 6 input images.

of Human-Object Interaction (HOI). HOI composition demands precise spatial relationships, realistic occlusions, and natural interactions between subjects and objects. Furthermore, current approaches either remain closed-source without disclosing critical implementation details, or they only support a limited number of input images and fixed resolutions during inference, such as Qwen-Image-Edit-2509.

To address these limitations, we present Skywork UniPic 3.0, a unified framework that seamlessly integrates text-to-image generation, single-image editing, and multi-image composition within a single model. Our key insight is to formulate all three tasks as conditional generation on a unified sequence representation, enabling a common architecture and training objective. This unified perspective not only simplifies model design but also facilitates knowledge transfer across tasks.

Our main contributions are threefold:

First, we propose a comprehensive data curation pipeline specifically tailored for multi-image composition. Recognizing that data quality outweighs quantity for this delicate task, we construct a high-quality dataset of 215k multi-image composition examples with a focus on challenging HOI scenarios. Our pipeline employs multi-stage filtering to ensure semantic coherence, visual compatibility, and composition quality, demonstrating that a carefully curated, moderately-sized dataset is sufficient to train a state-of-the-art model.

Second, we introduce a novel sequence modeling paradigm for multi-image composition. Specifically, we concatenate the noisy latent variables of the target output image with the latents of all reference images (which may be absent) along the sequence dimension to form a unified long sequence. This formulation enables our model to simultaneously train on text-to-image generation, single-image editing, and multi-image composition tasks while maintaining architectural simplicity. The unified sequence structure naturally accommodates variable numbers of input images and arbitrary output resolutions within a flexible pixel budget.

Third, we pioneer the integration of score-regularized continuous-time consistency models (rCM) into large-scale image generation frameworks. The distilled model produces high-fidelity results in merely 1–4 inference steps, achieving a remarkable 15× to 50× speedup over standard diffusion samplers, without sacrificing generation quality. We extensively evaluate Skywork UniPic 3.0 on established single-image editing benchmarks and our newly proposed MultiCom-Bench for multi-image composition. Our model achieves state-of-the-art performance on GEditBench-EN and ImgEdit-Bench, demonstrating the effectiveness of our unified framework. More importantly, on MultiCom-Bench, Skywork UniPic 3.0 surpasses recent strong baselines including Nano-Banana [4] and Seedream 4.0 [16]. These results validate that our careful data curation and novel training paradigm directly translate to superior composition quality.

In summary, this work presents the first systematic study of high-quality multi-image composition, provides a unified and efficient solution, and establishes a new state-of-the-art in the field. We believe our findings will benefit the community and inspire future research in unified generative frameworks.

2 Related Work

Image Foundation Models.

Image Generation with Few Steps. Efficient diffusion sampling is typically achieved by distribution matching or trajectory mapping. Distribution matching seeks to approximate the distribution of a teacher model using a few-step student model. Methods such as ADD [14] employ the Jensen-Shannon divergence to minimize the divergence of two distributions, while DMD [26, 25] use the reverse Kullback-Leibler divergence to learn the teacher distribution. These methods can produce compelling results but fail to capture full distribution due to the mode-seeking nature of the divergences they minimize. In contrast, trajectory mapping focuses on distilling the noise-to-data trajectory of a teacher model. Notable approaches include progressive distillation [13], consistency models [17, 6] and rectified distillation [8, 9]. sCM proposes a TrigFlow scheduler and a series techniques to simplify and stabilize continuous-time consistency training. Although trajectory mapping methods reduce the number of sampling steps, the generation quality still falls behind distribution distillation. Hybrid methods, such as DOLLAR [3] and rCM [27], combine both distribution matching and trajectory mapping for improved performance. We adopt the hybrid framework for few-step post-training, but propose a more principled formulation and a more efficient implementation.

3 Preliminary

Diffusion Models Given the noise from the Gaussian distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the clean data from the data distribution $\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})$, the diffusion models learn to map the noise distribution to the data distribution. Given the time range $t \in [0, T]$, the forward process utilizes the coefficients α_t and σ_t, such that $\boldsymbol{x}_t = \alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\varepsilon}$. UniPic 3.0 adopts the flow matching formulation, where $\alpha_t = 1 - t$ and $\sigma_t = t$ are in the range $t \in [0, 1]$. The Probability-Flow Ordinary Differential Equation (PF-ODE) of flow matching is: $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{\varepsilon} - \boldsymbol{x}$. The Training objective is given by:

$$\mathcal{L}_\theta^{\text{FM}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t} \left\| \boldsymbol{F}_\theta(\boldsymbol{x}_t, t) - \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} \right\|_2^2 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t} \left\| \boldsymbol{F}_\theta(\boldsymbol{x}_t, t) - (\boldsymbol{\varepsilon} - \boldsymbol{x}) \right\|_2^2, \tag{1}$$

where F_θ is the neural network with parameters θ. The sampling procedure begins with Gaussian noise $\boldsymbol{\varepsilon}$ and solves PF-ODE $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}$ from $t = 1$ to $t = 0$ with numerical solvers, usually taking multiple numbers of function evaluations.

Consistency Models A consistency model (CM) $\boldsymbol{f}_\theta(\boldsymbol{x}_t, t)$ learns directly the mapping from any point $\boldsymbol{x}_t$ on the trajectory to clean data $\boldsymbol{x}_0$. Conventionally, consistency models are parameterized as follows:

$$\boldsymbol{f}_\theta(\boldsymbol{x}_t, t) = c_{\text{skip}}(t)\boldsymbol{x}_t + c_{\text{out}}(t)\boldsymbol{F}_\theta(\boldsymbol{x}_t, t), \tag{2}$$

where $c_{\text{skip}}(t)$ and $c_{\text{out}}(t)$ are time-dependent coefficients that satisfy: $c_{\text{skip}}(0) = 1$ and $c_{\text{out}}(1) = 0$ to ensure the boundary condition $\boldsymbol{f}_\theta(\boldsymbol{x}, 0) \equiv \boldsymbol{x}$.

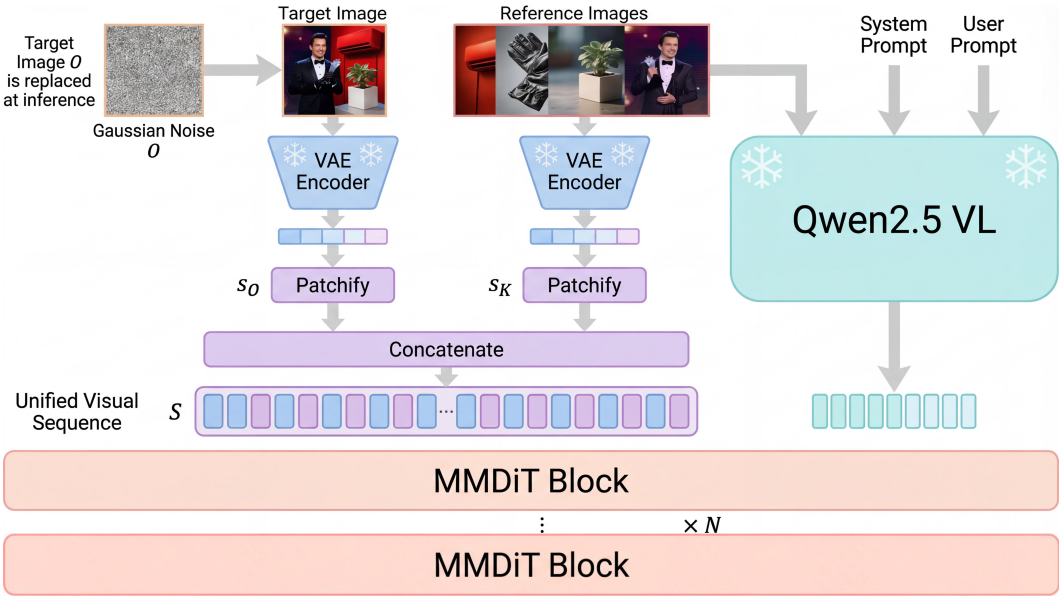

Figure 2: The overall pipeline of UniPic 3.0.

Learning the direct mapping to clean data is highly unstable and is lean to collapse. Therefore, CMs adopt a soft training objective to force the consistency property:

$$\mathcal{L}_\theta^{\mathrm{CM}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t}[w(t)d(\boldsymbol{f}_\theta(\boldsymbol{x}_t,t) - \boldsymbol{f}_{\theta^-}(\boldsymbol{x}_{t-\Delta t},t-\Delta t))], \tag{3}$$

where $\theta^- = \texttt{stopgrad}(\theta)$ means no gradients, $w(t)$ is the weighting function, Δt is the time interval, and $d(\cdot,\cdot)$ is the metric function, such as $\mathcal{L}_2$ loss and $\mathcal{L}_1$ loss.

Eqn. (3) is the optimization objective for discrete-time CMs, which suffers discretization errors induced by the time interval Δt. To reduce this error, continuous-time CMs adopt the limiting case: $\Delta t \to 0$. When choosing $d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$, the gradient of continuous-time CM can be simplified as follows:

$$\nabla_\theta \lim_{\Delta t \to 0} \frac{\mathcal{L}_\theta^{\mathrm{CM}}}{\Delta t} = \nabla_\theta \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t}\left[w(t)\boldsymbol{f}_\theta^\top(\boldsymbol{x}_t,t)\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}(\boldsymbol{x}_t,t)}{\mathrm{d}t}\right], \tag{4}$$

where $\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}(\boldsymbol{x}_t,t)}{\mathrm{d}t}$ is the tangent of $\boldsymbol{f}_{\theta^-}$ along the trajectory of the PF-ODE. Specifically, sCM proposes a TrigFlow transport for continuous-time consistency training, where $\alpha_t = \cos(t)$, $\sigma_t = \sin(t)$, $c_{\text{skip}} = \cos(t)$ and $c_{\text{out}}(t) = -\sin(t)$.

Distribution Matching Distillation Unlike consistency models, which learn to map the PF-ODE trajectory, distribution matching methods aim to match the student generation distribution p_θ to the teacher generation distribution p_{teacher}. In this case, samples are generated in a few-steps $\boldsymbol{x} \sim p_\theta$ via $\boldsymbol{x} = \boldsymbol{G}_\theta(\boldsymbol{\varepsilon})$[1], where $\boldsymbol{\varepsilon}$ represents Gaussian noise. To minimize the difference between p_θ and p_{teacher}, the f-divergence [12, 22] is used as an optimization objective:

$$\min_\theta[\mathcal{D}_f(p_\theta \parallel p_{\text{teacher}})] = \min_\theta\left[\int_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})} p_{\text{teacher}}(\boldsymbol{x}) f\left(\frac{p_\theta(\boldsymbol{x})}{p_{\text{teacher}}(\boldsymbol{x})}\right)\mathrm{d}\boldsymbol{x}\right]. \tag{5}$$

The choice of function $f(\cdot)$ determines the type of divergences, such as the reverse Kullback-Leibler (KL) divergence [26, 10], the Fisher divergence [28] and the Jensen-Shannon divergence [15, 14].

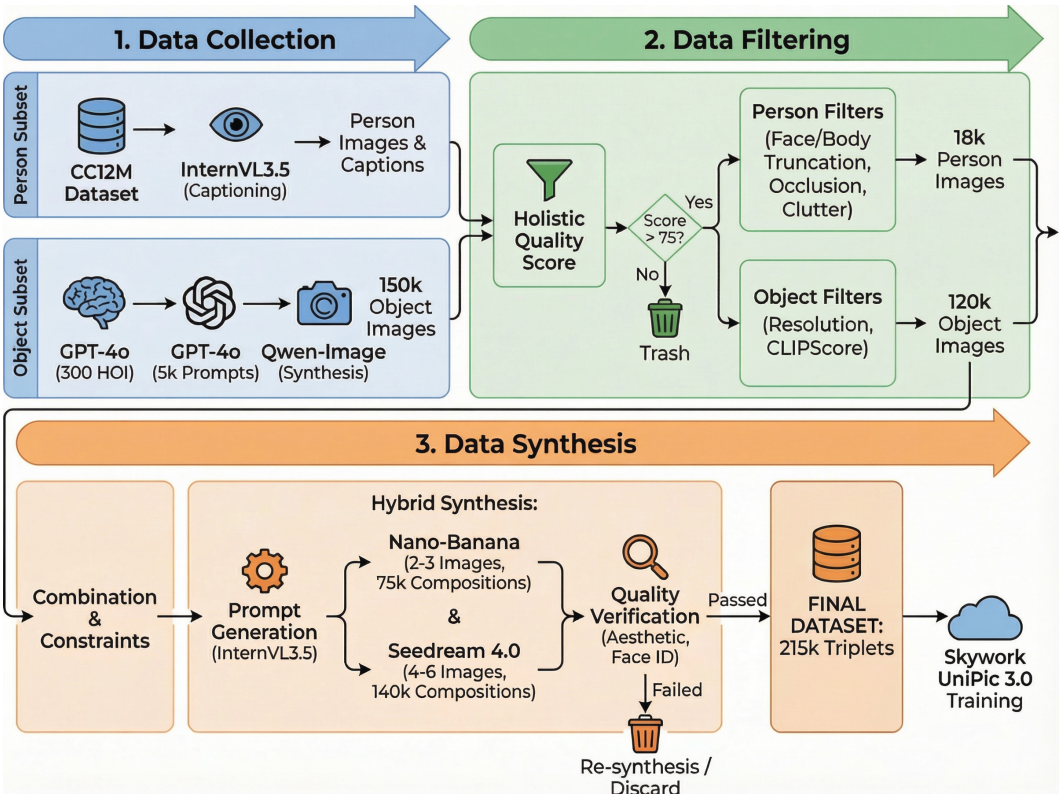

Figure 3: The overall data curation pipeline of UniPic 3.0.

4 Method

4.1 Data Curation

To achieve remarkable multi-image composition quality for Human-Object Interaction (HOI) tasks, we meticulously design a comprehensive data curation pipeline comprising three synergistic stages: data collection, data filtering, and data synthesis. This pipeline yields 215k high-quality (source images, instruction, target image) triplets that serve as the foundation for training Skywork UniPic 3.0.

Data Collection. We adopt a bifurcated strategy to source diverse and composition-ready person and object images. For the person subset, we curate human-centric images from the CC12M dataset, a large-scale corpus known for its rich variety of human poses, appearances, and real-world contexts. We employ InternVL3.5-38B [18] to generate dense, structurally-aware captions that explicitly describe human attributes, clothing, pose, and scene context—information crucial for subsequent HOI composition. For the object subset, we first prompt GPT-4o to generate 300 fine-grained object categories that are semantically compatible with human interaction (e.g., apparel items, handheld tools, musical instruments, furniture, sports equipment). For each category, GPT-4o then produces 5,000 diverse textual prompts emphasizing visual attributes, materials, and typical usage contexts. These prompts are fed into Qwen-Image [21] to synthesize 150k object images. This deliberate separation ensures balanced coverage across interaction types while maintaining photorealistic quality.

Data Filtering. We implement a rigorous multi-stage filtering protocol to ensure source image quality and composition suitability. First, we deploy InternVL3.5-38B [18] to assign a holistic quality score (0–100) to each image based on resolution, sharpness, aesthetic appeal, and semantic clarity, discarding images scoring below 75. For person images, we apply specialized face and body detectors to filter out cases with truncated heads (<90% face visibility), occluded primary subjects

[1] $\boldsymbol{G}_\theta(\boldsymbol{\varepsilon}) = \Psi(\boldsymbol{F}_\theta, \boldsymbol{\varepsilon}, N)$, where Ψ is an ODE solver, $\boldsymbol{F}_\theta$ is the neural network and N is the number of function evaluations. This represents the process from pure Gaussian noise $\boldsymbol{\varepsilon}$ to generated samples.

(<60% foreground occupancy), or cluttered backgrounds (background complexity score > 0.7). For object images, we enforce minimum resolution requirements (> 768×768 pixels) and cross-check prompt-image alignment using CLIPScore, rejecting pairs with low similarity. This stringent curation retains only 18k person images and 120k object images, ensuring a clean and composition-ready inventory.

Data Synthesis. The synthesis stage carefully constructs valid source image combinations and generates corresponding target compositions. We sample 2–6 source images per composition while enforcing hard HOI compatibility constraints through a manually curated conflict matrix: for instance, a person cannot simultaneously wear two pairs of footwear or interact with two musical instruments in the same hand. This prevents physically implausible scenes. For each valid combination, we prompt InternVL3.5-38B [18] to generate a cohesive composition prompt that describes natural spatial arrangements, realistic occlusion relationships, and harmonious scene lighting—a critical step that bridges the semantic gap between disparate sources. Empirically, we observe that Nano-Banana [4] exhibits degraded facial identity preservation when handling >3 input images, whereas Seedream 4.0 [16] maintains superior consistency across 4–6 images. We therefore adopt a hybrid synthesis strategy: Nano-Banana generates 75k compositions for 2–3 image subsets, while Seedream 4.0 [16] produces 140k compositions for 4–6 image subsets. Each synthesized target undergoes automatic quality verification via Aesthetic Score and face identity preservation checks, with failed cases re-synthesized or discarded.

MultiCom-Bench. Recognizing the absence of standardized evaluation protocols for multi-image composition, we construct MultiCom-Bench, a carefully curated benchmark comprising 200 high-quality triplets specifically targeting HOI scenarios. The benchmark is balanced across interaction types and input complexity (100 triplets with 2–3 source images, 100 with 4–6 images). Following VIEScore [7], we have designed stable and effective evaluation templates that assess model-generated results across multiple dimensions, including adherence to composition instructions, image quality, and facial consistency. This benchmark will be released to facilitate future research in multi-image composition.

4.2 Training Paradigm

In this section, we describe the training paradigm that enables our model to perform multi-image fusion by casting it as conditional generation over a unified visual sequence. Our model architecture follows that of Qwen-Image [21], which incorporates Qwen2.5-VL [2] as the condition encoder, employs a VAE as the image tokenizer, and utilizes MMDiT as the backbone diffusion model.

Latent Encoding. For each training instance, we sample a target image O and a set of reference images $\{I_1, \ldots, I_K\}$. Each image is encoded into a latent tensor using the VAE encoder:

$$z_O = f_{\text{vae-enc}}(O), \quad z_k = f_{\text{vae-enc}}(I_k),\ k = 1, \ldots, K, \tag{6}$$

where $z_O, z_k \in \mathbb{R}^{1\times C\times H'\times W'}$, C is the latent channel dimension, and (H', W') is the downsampled spatial resolution.

Patch-wise Packing. Following the design of Qwen-Image, each latent tensor is reshaped into a sequence of patches by a deterministic packing operation:

$$s_O = \text{pack}(z_O) \in \mathbb{R}^{N_O\times D}, \quad s_k = \text{pack}(z_k) \in \mathbb{R}^{N_k\times D}, \tag{7}$$

where pack$(\cdot)$ rearranges 2×2 spatial neighborhoods into tokens, yielding a sequence length $N_\cdot$ and feature dimension D. This operation is invertible given the spatial metadata.

Unified Visual Sequence. We construct a single unified latent sequence by concatenating the packed target and reference latents along the sequence dimension:

$$S = [s_O \,\|\, s_1 \,\|\, \ldots \,\|\, s_K] \in \mathbb{R}^{N_{\text{tot}}\times D}, \tag{8}$$

where

$$N_{\text{tot}} = N_O + \sum_{k=1}^{K} N_k. \tag{9}$$

In addition, we maintain a set of shape descriptors

$$\mathcal{H} = \{h_O, h_1, \ldots, h_K\}, \tag{10}$$

where each $h_\cdot$ encodes the latent height and width of the corresponding image. These descriptors are passed into the transformer to preserve spatial structure and enable exact unpacking during reconstruction.

At inference time, we replace the true target latents with pure Gaussian noise while keeping reference latents.

4.3 Post-training for Few-step Generation

Trajectory mapping methods, such as consistency models, can effectively reduce sampling steps and reserve generation diversity but suffer from fine-grained generation quality. In contrast, distribution matching methods, such as GAN and DMD, can produce more compelling results but fail to capture full distribution due to the mode-seeking nature of the divergences they minimize.

Our post-training framework integrates principles from DOLLAR and rCM to transform a multi-step DiT model to a few-step alternative. We propose a hybrid framework that combines trajectory mapping and distribution matching, ensuring few-step generation with both high fidelity and diversity.

Continuous-time Consistency Training for Flow Matching. sCM simplifies and stabilizes continuous-time CMs using the TrigFlow formulation. To enable continuous-time consistency training for flow matching (FM) models, SANA-Sprint and rCM transform a pre-trained flow matching model into a TrigFlow model through mathematical input and output transformations. However, this approach changes the time variables from a standard linear schedule to a trigonometric schedule, leading to numerical instability and precision errors. The input and output transformations induce extra gradient terms, which affects the gradient variance and makes the training unstable.

To address these challenges, we adopt a more principled consistency flow matching training for UniPic 3.0 model. For linear transport in flow matching, the consistency models are parameterized as: $\boldsymbol{f}_\theta(\boldsymbol{x}_t, t) = \boldsymbol{x}_t - t\boldsymbol{F}_\theta(\boldsymbol{x}_t, t)$, leading to the gradient:

$$\nabla_\theta \boldsymbol{f}_\theta(\boldsymbol{x}_t, t) = -t\nabla_\theta \boldsymbol{F}_\theta(\boldsymbol{x}_t, t). \tag{11}$$

The tangent of $\boldsymbol{f}_{\theta^-}$ has the explicit form:

$$\frac{\mathrm{d}\boldsymbol{f}_{\theta^-}(\boldsymbol{x}_t, t)}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} - \boldsymbol{F}_{\theta^-}(\boldsymbol{x}_t, t) - t\frac{\mathrm{d}\boldsymbol{F}_{\theta^-}(\boldsymbol{x}_t, t)}{\mathrm{d}t} = \boldsymbol{\varepsilon} - \boldsymbol{x} - \boldsymbol{F}_{\theta^-}(\boldsymbol{x}_t, t) - t\frac{\mathrm{d}\boldsymbol{F}_{\theta^-}(\boldsymbol{x}_t, t)}{\mathrm{d}t}. \tag{12}$$

Taking the explicit form of Eqn. (11) and Eqn. (12) into Eqn. (4), the gradient is:

$$\nabla_\theta \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon},t}\left[\boldsymbol{F}_\theta^\top(\boldsymbol{x}_t, t) \cdot \left(-w(t) \cdot t \cdot \frac{\mathrm{d}\boldsymbol{f}_{\theta^-}(\boldsymbol{x}_t, t)}{\mathrm{d}t}\right)\right]. \tag{13}$$

Taking the identity $\nabla_\theta \mathbb{E}[\boldsymbol{F}_\theta^\top \boldsymbol{y}] = \frac{1}{2}\nabla_\theta \mathbb{E}[\|\boldsymbol{F}_\theta - \boldsymbol{F}_{\theta^-} + \boldsymbol{y}\|_2^2]$[2] and $w(t) = \frac{1}{t}$, the consistency loss is defined as:

$$\mathcal{L}_\theta^{\mathrm{CM}} = \left\| F_\theta(\boldsymbol{x}_t, t) - \left(\boldsymbol{\varepsilon} - \boldsymbol{x} - t\frac{\mathrm{d}\boldsymbol{F}_{\theta^-}(\boldsymbol{x}_t, t)}{\mathrm{d}t}\right)\right\|_2^2, \tag{14}$$

where we calculate $\frac{\mathrm{d}\boldsymbol{F}_{\theta^-}(\boldsymbol{x}_t, t)}{\mathrm{d}t} = \frac{1}{2\epsilon}(\boldsymbol{F}_{\theta^-}(\boldsymbol{x}_{t+\epsilon}, t+\epsilon) - \boldsymbol{F}_{\theta^-}(\boldsymbol{x}_{t-\epsilon}, t-\epsilon))$ with the finite-difference approach [27, 19], where we set $\epsilon = 5 \times 10^{-3}$ by default.

Distribution Matching Distillation. We adopt the reverse KL divergence to match the student distribution with the teacher distribution, leading to the gradient of Eqn. (5) as:

$$\begin{aligned}\nabla_\theta \boldsymbol{D}_{\mathrm{KL}}(p_\theta \parallel p_{\text{teacher}}) &= \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{x},t}[-w(t)(\nabla_{\boldsymbol{x}_t}\log p_{\text{teacher}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t}\log p_{\theta^-}(\boldsymbol{x}_t))\nabla_\theta \boldsymbol{G}_\theta(\boldsymbol{\varepsilon})] \\ &\approx \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{x},t}[-w(t)(\nabla_{\boldsymbol{x}_t}\log p_{\text{teacher}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t}\log p_\phi(\boldsymbol{x}_t))\nabla_\theta \boldsymbol{G}_\theta(\boldsymbol{\varepsilon})],\end{aligned} \tag{15}$$

where $\nabla_{\boldsymbol{x}_t}\log p_{\text{teacher}}(\boldsymbol{x}_t)$ and $\nabla_{\boldsymbol{x}_t}\log p_{\theta^-}(\boldsymbol{x}_t)$ represent the score functions for the teacher and student distributions. And ϕ represents the parameters of a fake score network that models the

[2]This identity is valid for any arbitrary vector $\boldsymbol{y}$, provided that $\boldsymbol{y}$ is independent of the parameter θ.

distribution of the students, as the student score score $\nabla_{\boldsymbol{x}_t} \log p_{\theta^-}(\boldsymbol{x}_t)$ is not intractable for the few-step generator F_θ. Typically, the fake score network F_ϕ is initialized from the teacher model $\boldsymbol{F}_{\text{teacher}}$ added with LoRA, and only the LoRA parameters are tunable during training. It adopts the flow matching loss as in Eqn. (1) with data from the student model $\boldsymbol{x} \sim p_\theta$, thus serving as a proxy for the student distribution.

The gradient term in Eqn. (15) is:

$$\begin{aligned}\boldsymbol{g} &= -(\nabla_{\boldsymbol{x}_t} \log p_{\text{teacher}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{x}_t)) \\ &= -\left(-\frac{\boldsymbol{x}_t + (1-t)\boldsymbol{F}_{\text{teacher}}(\boldsymbol{x}_t, t)}{t} - \left(-\frac{\boldsymbol{x}_t + (1-t)\boldsymbol{F}_\phi(\boldsymbol{x}_t, t)}{t}\right)\right) \\ &= \frac{1-t}{t}(\boldsymbol{F}_{\text{teacher}}(\boldsymbol{x}_t, t) - \boldsymbol{F}_\phi(\boldsymbol{x}_t, t)),\end{aligned} \tag{16}$$

where $\boldsymbol{x}_t = (1-t)\boldsymbol{x} + t\hat{\boldsymbol{\varepsilon}}$ with independently sampled noise $\hat{\boldsymbol{\varepsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ To reduce the variance of the gradient, we adopt the weighting function $w(t) = \frac{t}{1-t} \cdot \frac{H \cdot W \cdot C}{\|\boldsymbol{x} - \hat{\boldsymbol{\varepsilon}} + \boldsymbol{F}_\phi(\boldsymbol{x}_t, t)\|_1}$, where H, W, and C represent the height, width, and channel of $\boldsymbol{x}$. Therefore, the weighted gradient term is as follows:

$$w(t)\boldsymbol{g} = \frac{H \cdot W \cdot C \cdot (\boldsymbol{F}_{\text{teacher}}(\boldsymbol{x}_t, t) - \boldsymbol{F}_\phi(\boldsymbol{x}_t, t))}{\|\boldsymbol{x} - \hat{\boldsymbol{\varepsilon}} + \boldsymbol{F}_\phi(\boldsymbol{x}_t, t)\|_1}. \tag{17}$$

The distribution matching distillation loss is defined as:

$$\mathcal{L}_\theta^{\text{DMD}} = \frac{1}{2}\left\|\boldsymbol{G}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{G}_{\theta^-}(\boldsymbol{x}_t, t) + \frac{H \cdot W \cdot C \cdot (\boldsymbol{F}_{\text{teacher}}(\boldsymbol{x}_t, t) - \boldsymbol{F}_\phi(\boldsymbol{x}_t, t))}{\|\boldsymbol{x} - \hat{\boldsymbol{\varepsilon}} + \boldsymbol{F}_\phi(\boldsymbol{x}_t, t)\|_1}\right\|_2^2. \tag{18}$$

Final Objective. Our framework culminates in a comprehensive framework for few-step learning. We train the network $\boldsymbol{F}_\theta$ with both consistency loss Eqn. (14) and distribution matching loss Eqn. (18), leading to the final objective as follows:

$$\mathcal{L}_\theta = \lambda_{\text{CM}}\mathcal{L}_\theta^{\text{CM}} + \lambda_{\text{DMD}}\mathcal{L}_\theta^{\text{DMD}}, \tag{19}$$

where λ_{CM} and λ_{DMD} are the loss weightings for the two losses, respectively.

5 Experiments

5.1 Setup

The training data comprises 338k multi-image composition samples (including 215k internally constructed samples and Mico-150K [1]) and 381k single-image editing samples from open-source datasets (Nano-consistent-150K [23], Pico-Banana-400K [11]). We perform full-parameter fine-tuning of the DiT model, training it for $100K$ steps with a global batch size of 64 and a learning rate of 1×10^{-4}. The training employs cosine learning rate annealing and the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1 \times 10^{-8}$, and a weight decay of 0.05.

5.2 Main results.

Tab. **??** summarizes our main comparative results, evaluating our models against other advanced methods on image editing, and multi-image composition. We compare with GPT-4o [5], Nano-Banana [4], Seedream 4.0 [16]. The results clearly demonstrate the exceptional performance of our approach across major benchmarks, highlighting its powerful unified capabilities. Detailed comparisons for image editing and multi-image composition are provided in subsequent subsections. We further present ablation studies in section 5.3.

Image Editing. We first evaluate UniPic 3.0 on single-image editing using two standard benchmarks: ImgEdit-Bench [24] and GEdit-Bench [7]. As shown in Table 1, UniPic 3.0 achieves overall scores of 4.35 on ImgEdit-Bench and 7.55 on GEdit-Bench. These results indicate that UniPic 3.0 preserves strong single-image editing performance while being equipped with unified multi-image composition capabilities.

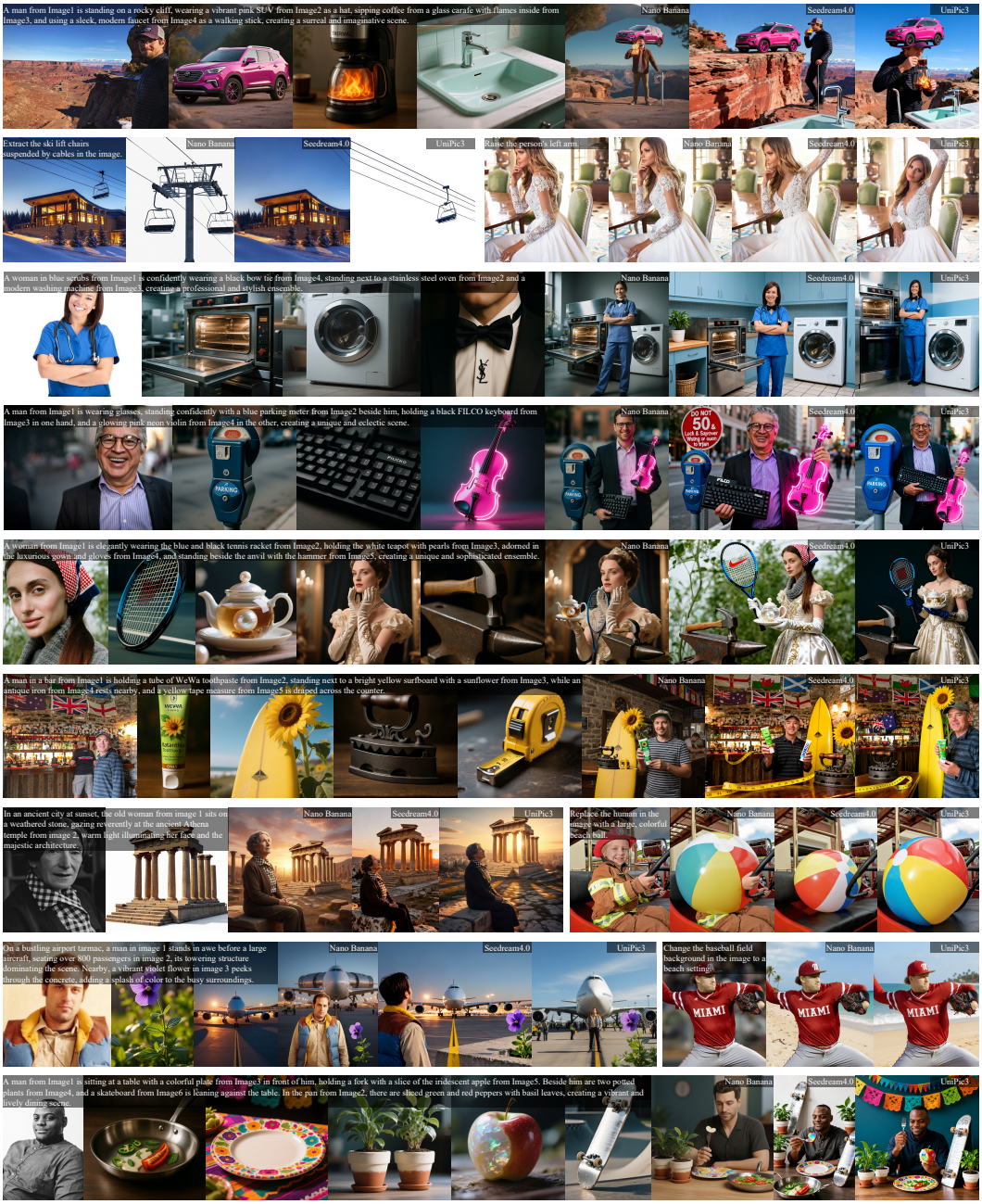

Figure 4: Qualitative comparison among Nano-Banana, SeedDream4, and UniPic3. Our model demonstrates competitive or even superior performance in instruction-guided image editing and composition.

Table 1: Performance on ImgEdit Bench and GEdit Bench.

Model	ImgEdit-Bench ↑										GEdit-Bench ↑		
	Extract	Style	Background	Add	Remove	Replace	Adjust	Compose	Action	Overall	G_SC	G_PQ	G_O
Qwen-Image-Edit [21]	3.47	4.80	4.32	4.26	3.87	4.58	4.45	3.91	4.59	4.25	8.18	7.87	7.68
Qwen-Image-Edit-2509 [21]	3.51	4.84	4.36	4.43	4.29	4.66	4.42	3.72	4.58	4.31	8.12	8.01	7.61
Nano Banana [4]	3.89	4.2	4.32	4.33	4.39	4.55	4.36	3.42	4.48	4.22	7.43	8.14	7.20
Seedream 4.0 [16]	2.96	4.76	4.22	4.47	4.25	4.42	4.31	3.11	4.45	4.11	8.24	7.86	7.66
UniPic 2.0 [20]	1.86	4.53	4.73	4.48	4.00	4.73	4.18	3.82	4.22	4.06	7.63	7.17	7.10
UniPic 3.0	3.31	4.97	4.35	4.45	4.46	4.71	4.44	3.77	4.69	4.35	8.12	7.79	7.55

Table 2: Performance comparison of different models on multi-image composition.

Model	MultiCom-Bench		
	2-3 imgs	**4-6 imgs**	**Overall**
Nano-Banana	0.7982	0.6466	0.7224
Seedream 4.0	0.7997	0.6197	0.7088
Ours	0.8214	0.6296	**0.7255**

Multi-image Composition.

5.3 Ablation Studies

6 Conclusions

This work presents Skywork UniPic 3.0.

7 Contributors

Core contributors: Hongyang Wei*, Baixin Xu*, Hongbo Liu*, Cyrus Wu†, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yang Liu‡, Xuchen Song‡, Eric Li‡

Contributors: Yidan Xietian, Chuanxin Tang, Zidong Wang, Yichen Wei, Liang Hu, Boyi Jiang, William Li, Ying He, Yahui Zhou

* Equal contribution.
† Project Lead.
‡ Corresponding author.

References

[1] A113N-W3I. Mico-a comprehensive dataset and benchmark advancing multi-images composition. GitHub repository, 2025. 8

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025. 6

[3] Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17961–17971, 2025. 3

[4] Google. Nano banana, 2025. Accessed: 2025. 1, 3, 6, 8, 9

[5] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024. 8

[6] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. arXiv preprint arXiv:2310.02279, 2023. 3

[7] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025. 6, 8

[8] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In ICLR, 2023. 3

[9] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In The Twelfth International Conference on Learning Representations, 2023. 3

[10] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. Advances in Neural Information Processing Systems, 36:76525–76546, 2023. 4

[11] Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing, 2025. 8

[12] Alfréd Rényi. On measures of entropy and information. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics, volume 4, pages 547–562. University of California Press, 1961. 4

[13] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512, 2022. 3

[14] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In SIGGRAPH Asia 2024 Conference Papers, pages 1–11, 2024. 3, 4

[15] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In European Conference on Computer Vision, pages 87–103. Springer, 2024. 4

[16] Team Seedream, :, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. Seedream 4.0: Toward next-generation multimodal image generation, 2025. 1, 3, 6, 8, 9

[17] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In International Conference on Machine Learning, pages 32211–32252. PMLR, 2023. 3

[18] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025. 5, 6

[19] Zidong Wang, Yiyuan Zhang, Xiaoyu Yue, Xiangyu Yue, Yangguang Li, Wanli Ouyang, and Lei Bai. Transition models: Rethinking the generative learning objective. arXiv preprint arXiv:2509.04394, 2025. 7

[20] Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. arXiv preprint arXiv:2509.04548, 2025. 9

[21] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025. 5, 6, 9

[22] Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. arXiv preprint arXiv:2502.15681, 2025. 4

[23] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. https://arxiv.org/abs/2508.09987, 2025. 8

[24] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275, 2025. 8

[25] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. Advances in Neural Information Processing Systems, 37:47455–47487, 2025. 3

[26] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6613–6623, 2024. 3, 4

[27] Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. arXiv preprint arXiv:2510.08431, 2025. 3, 7

[28] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In Forty-first International Conference on Machine Learning, 2024. 4